SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100293

Le test de substituabilité à l'épreuve des corpus : utiliser l'analyse distributionnelle automatique pour l'étude des relations lexicales

François Morlane-Hondère & Cécile Fabre

CLLE-ERSS, Université de Toulouse
{francois.morlane,cecile.fabre}@univ-tlse2.fr

1 Introduction

Dans « Distributional structure », Harris (1954) formule l'*hypothèse distributionnelle*, selon laquelle la proximité sémantique de deux mots se traduit par une similarité au niveau de leurs contextes d'apparition respectifs (leurs *distributions*). Ce principe de base a été automatisé dès le début des années 1990, en particulier par Hindle (1990), Hearst (1992), Ruge (1992) ou Grefenstette (1994). Il a été dans un premier temps implémenté dans des systèmes orientés vers la création de thesaurus à partir de textes spécialisés, puisque c'est précisément pour traiter des données de ce type qu'a été formulée l'hypothèse distributionnelle : les restrictions sélectionnelles particulièrement fortes qui régissent la distribution des mots dans ces textes en font un matériau propice à la délimitation de classes distributionnelles. On peut ainsi citer les travaux de Grefenstette (1992) portant sur des textes issus du domaine médical, ou, parmi les études réalisées à partir de textes en français sur le même domaine, les travaux de Nazarenko *et al.* (1997) et Bouaud *et al.* (2000).

L'accroissement de la quantité de textes accessibles au format électronique a permis par la suite au modèle distributionnel de gagner en popularité en entrainant l'extension de son domaine d'application aux textes non spécialisés comme les textes journalistiques (Dias *et al*. 2010) ou le Web (Turney 2008 ; Agirre *et al.* 2009). Le fait d'appliquer l'analyse distributionnelle (AD) à des textes ne relevant pas de domaines de spécialité produit toutefois des résultats moins faciles à caractériser. On rencontre une réelle difficulté à évaluer les sorties de ces programmes, qui produisent, appliqués à de vastes corpus, des résultats pléthoriques. La prédominance de relations floues, non recensées dans les ressources lexicales traditionnellement utilisées en TAL rend de telles données difficiles à évaluer selon des procédures classiques ; leur apport dans des applications comme la recherche d'information n'a pas été démontré (Van der Plas 2008 ; Sahlgren 2006). Plusieurs travaux ont montré néanmoins que ces techniques pouvaient permettre de repérer des relations sémantiques de différents types – synonymie, généricité, analogie, etc. (Turney 2008 ; Baroni et Lenci 2011).

Notre objectif, dans le cadre de cet article, est d'étudier les résultats produits par un système d'analyse distributionnelle automatique afin de mieux comprendre sous quelles conditions le critère distributionnel permet de repérer les relations lexicales les plus usuelles – synonymie, antonymie, hyperonymie, méronymie. Le test de substituabilité est le critère clé auquel les lexicologues ont recours pour identifier la plupart des relations de nature paradigmatique entre mots (Cruse 1986 ; Murphy 2003). Un système d'analyse distributionnelle automatique offre précisément la possibilité de mettre en œuvre ce test à grande échelle, sur un large corpus. Il constitue un outil intéressant pour la vérification empirique de ce principe et, de façon plus générale, pour l'étude de ces relations sémantiques en corpus. Nous avons choisi d'aborder cette question en confrontant les résultats du programme d'AD dont nous disposons avec des données issues de ressources lexicales recensant différents types de relations sémantiques (synonymie, antonymie, hyperonymie, méronymie). Cette confrontation montre de forts décalages entre la ressource distributionnelle et ces lexiques. Si une part importante des paires reliées dans les lexiques sont des voisins distributionnels, beaucoup d'entre elles ne sont pas identifiées par l'AD, même quand il s'agit d'unités lexicales fréquentes dans le corpus. Nous essayons de comprendre les raisons de ces

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100293

décalages en nous appuyant sur les informations que nous fournit l'analyse automatique. Cette étude est menée sur des données en français.

Dans un premier temps, nous décrivons la méthode qui a été mise en œuvre pour obtenir la base distributionnelle sur laquelle nous appuyons nos analyses (2.1), que nous comparons ensuite à deux ressources externes, à savoir le *Dictionnaire Électronique des Synonymes* et le réseau *JeuxDeMots* (2.2). Après avoir mesuré l'intersection de ces deux ressources et de notre base de voisins (3.1), nous comparons, en termes de propriétés générales, les couples de voisins qui apparaissent dans les ressources et ceux qui n'y apparaissent pas (3.2). Ces deux démarches relèvent d'une approche *quantitative*, par opposition aux approches *qualitatives* que nous adoptons dans les sections qui suivent. Nous y analysons d'abord les raisons pour lesquelles des paires liées par des relations sémantiques identifiées dans les lexiques ne figurent pas parmi les voisins, à travers l'observation de couples de synonymes, d'antonymes, d'hyperonymes et de méronymes (3.3). Nous utilisons enfin (3.4) les mesures de précision et de rappel pour centrer notre analyse sur les mots dont les synonymes ne sont quasiment pas repérés par l'AD.

2 Présentation des données

Cette étude repose sur la confrontation des résultats de l'analyse distributionnelle automatique et de deux ressources lexicales – le *Dictionnaire Électronique des Synonymes* du CRISCO et une partie du réseau de la base JeuxDeMots du LIRMM. Nous présentons chaque ressource en commençant par la ressource distributionnelle.

2.1 Les voisins distributionnels

La base distributionnelle utilisée dans cette étude a été obtenue à partir du traitement d'un corpus constitué de l'intégralité des articles de l'encyclopédie en ligne Wikipédia dans une version datant d'avril 2007. Dans la suite de l'article, cette base est désignée sous le nom *Voisins de Wikipédia*[1] (VDW). Le corpus utilisé compte environ 194 millions de mots. Ce choix est motivé par des considérations pratiques de disponibilité de la ressource, mais également par l'intérêt présenté par une collection de textes homogènes du point de vue du genre mais variés sur le plan thématique, ce qui permet d'observer le comportement d'unités lexicales sémantiquement très diverses. Le modèle distributionnel qui a été appliqué a été conçu par Didier Bourigault à partir des sorties de l'analyseur Syntex (Bourigault 2002 et 2007). Il s'agit donc d'un modèle d'analyse distributionnelle « structuré » (Baroni et Lenci 2010) : le contexte de chaque mot est composé de l'ensemble des mots qui entretiennent avec lui une fonction syntaxique dans la phrase. Cette caractéristique fournit des éléments plus précis pour l'interprétation des résultats que la mise en œuvre de simples cooccurrences.

La procédure d'analyse a été exposée dans (Bourigault 2002). Nous en décrivons ici les principaux aspects. L'analyseur Syntex modélise les dépendances entre les mots d'une phrase sous la forme de triplets de lemmes <gouverneur, relation, dépendant>. Seuls les triplets constitués de noms (ou de syntagmes nominaux), de verbes et d'adjectifs sont pris en compte pour le calcul des voisins. Les relations syntaxiques considérées sont les relations sujet, objet, la modification adjectivale, ainsi que les relations prépositionnelles, celles-ci étant décrites par le biais de la préposition impliquée. On obtient ainsi les triplets suivants après analyse de la phrase *Le navajo utilise un système de numérotation décimal* :

<utiliser, SUJ, navajo>
<utiliser, OBJ, système de numérotation>
<utiliser, OBJ, système>
<décimal, MOD, système de numérotation>
<décimal, MOD, système>
<système, de, numérotation>

L'analyse distributionnelle effectuée ensuite classe les mots selon un double rapprochement :

- les dépendants sont rapprochés sur la base des contextes gouverneur_relation qu'ils partagent. Ainsi, *système* est rapproché de *modèle* parce qu'ils sont objets des mêmes verbes (*calquer*, *mettre au point*, *imaginer*…). On parle alors de rapprochement entre arguments.
- les gouverneurs, munis de la relation, sont rapprochés sur la base des dépendants qu'ils régissent. Ainsi, *utiliser*_OBJ est rapproché de *posséder*_OBJ parce que les dépendants en position objet de l'un et de l'autre se recouvrent largement (ils partagent 780 lemmes différents : *particularité, ordinateur, tête*…). On parle alors de rapprochement entre prédicats.

Ces rapprochements sont calculés à l'aide d'une mesure de similarité entre les vecteurs de contextes associés aux mots, la mesure de Lin (1998). Le score de similarité de deux prédicats/arguments varie – de 0 à 1 – en fonction de plusieurs facteurs : le nombre de contextes partagés, le nombre de triplets différents dans lesquels chacun de deux mots apparaît (indice de productivité), le degré de spécificité du contexte qui permet d'effectuer le rapprochement. Ce dernier indice est calculé en utilisant la mesure d'information mutuelle : ainsi, le contexte *sentier_de*, qui permet de rapprocher *promenade* et *randonnée*, est plus informatif que le contexte *faire_OBJ*, qui a une distribution beaucoup plus étendue. Le premier a donc plus de poids que le second dans le calcul de similarité.

La valeur d'un ensemble de paramètres (score de la mesure de similarité, types de contextes considérés, seuil de fréquences des mots et des contextes, etc.) peut être ajustée ; ces choix ont nécessairement des conséquences sur les résultats (Van der Plas 2008 ; Baroni et Lenci 2011). Dans la version de la base que nous avons utilisée, nous avons opté pour les seuils suivants : les triplets pris en compte dans le calcul ont une fréquence supérieure ou égale à 5 ; le seuil de productivité a été également fixé à 5 ; le score de Lin considéré est supérieur ou égal à 0,1. La taille de la base obtenue avec ces réglages est de près de 4 millions de paires (3 922 657 exactement).

Les tableaux 1 et 2 donnent quelques illustrations de cette relation de voisinage. Le tableau 1 montre les premiers voisins du nom *expédition* (productivité de 248 : il apparait dans 248 triplets différents) en position de dépendant, par ordre décroissant de la mesure du Lin. Le nombre de contextes différents dans lesquels apparaît chaque lemme est indiqué dans la troisième colonne. Le deuxième tableau montre les voisins du verbe *réparer* lorsque l'on considère la position objet. Ce verbe comprend lui-même 45 cooccurrents dans cette position. La première ligne se lit de la manière suivante : *réparation* est voisin de *réparer*, car 14 mots apparaissent à la fois en position de complément du nom de *réparation*, et en position objet de *réparer*.

Catégorie	Lemme	Productivité	Nb. contextes partagés	Prox Lin
N	*campagne*	468	131	0,332
N	*mission*	470	130	0,322
N	*opération*	493	131	0,322
N	*voyage*	301	91	0,308
N	*croisade*	92	51	0,275

Tableau 1 : 5 voisins les plus fortement associés au nom *expédition* en position de dépendant.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100293

Catégorie	Lemme	Relation	Productivité	Nb. contextes partagés	Prox Lin
N	*réparation*	de	23	14	0,394
V	*endommager*	OBJ	48	15	0,257
N	*réfection*	de	11	7	0,216
V	*apercevoir*	OBJ	39	11	0,21
V	*démolir*	OBJ	25	9	0,209

Tableau 2 : 5 voisins les plus fortement associés au verbe *réparer* à travers la relation objet.

Ces exemples montrent la diversité des relations accessibles par le calcul distributionnel. Celui-ci détecte à la fois des relations lexicales de type synonymie et antonymie, des relations de dérivation, et des relations plus lâches (*tight* et *loose relations* selon Kilgarriff et Yallop 2000). Ainsi *apercevoir* est rapproché de *réparer* selon une relation très ténue : les deux verbes ont pour seul point commun de pouvoir s'appliquer à certains grands artéfacts (*clocher*, *vaisseau*, *navire*, etc.). La diversité des relations de similarité qui est détectée par l'AD n'est cependant pas l'objet de cet article. Nous nous focalisons ici sur la part de relations lexicales que repère l'AD. Ce point de vue est certes réducteur car cette évaluation externe de la ressource ne permet pas d'apprécier la qualité globale des relations sémantiques identifiées. En confrontant la ressource distributionnelle avec des lexiques existants, notre objectif est de mieux comprendre ce qui conditionne le repérage par l'AD des relations lexicales identifiées dans ces ressources.

2.2 Le Dictionnaire Électronique des Synonymes et JeuxDeMots

Le Dictionnaire Électronique des Synonymes du CRISCO[2] (Manguin *et al.* 2004), ou *DES*, est issu de la compilation des synonymes présents dans sept dictionnaires (dictionnaires analogiques et dictionnaires de synonymes). Il contient près de 400 000 couples de synonymes.

JeuxDeMots (JDM) est une ressource issue du *crowdsourcing* : elle est construite de façon collaborative par des locuteurs (experts et non-experts confondus) participant à un jeu en ligne[3] consistant à proposer une série de mots pour un mot-cible et une relation donnés (Lafourcade 2007). Les relations proposées incluent la synonymie, l'antonymie, l'hyperonymie et la méronymie, ainsi que des relations moins classiques comme les relations chose/lieu, agent/action, action/instrument, etc.

Afin de procéder à la mesure du taux de recouvrement entre les voisins de Wikipédia et les deux lexiques que nous avons choisis comme étalons, nous avons harmonisé les données pour assurer leur comparabilité :

- nous avons supprimé la mention de la relation associée aux prédicats de la base de voisins, pour ne conserver que le lemme. Par exemple, les verbes *arriver/venir* sont en relation dans plusieurs couples de voisins, car ils sont unis par la relation sujet ainsi que par différentes relations prépositionnelles (*à, en, avec*, etc.). Ces doublons ont été effacés pour ne retenir que la relation générique *Est-voisin*(*arriver*,*venir*).
- de la même façon, les couples de JDM sont également dédoublonnés dans la première phase de comparaison globale : lorsqu'un couple de mots est listé plusieurs fois par le biais de plusieurs relations sémantiques, nous ne considérons qu'une instance du couple (à titre d'exemple, le couple *montagne/sommet* apparaît à 9 reprises *via* les relations idée associée, chose/lieu, synonymie, méronymie, etc.).
- les relations de JDM ont été symétrisées (elles le sont par défaut dans le cas des deux autres ressources), c'est-à-dire que la relation A/B est systématiquement complétée par la relation B/A. Dans le cas des relations non symétriques que sont l'hyperonymie et la méronymie, l'orientation de la relation n'est donc pas considérée.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100293

- les couples de mots impliquant au moins une unité polylexicale ont été retirés des trois bases. Cette décision a été prise pour simplifier la procédure de comparaison. En particulier, Syntex lemmatise systématiquement chacun des éléments des unités polylexicales (*affaires étrangères* devient *affaire étranger*), ce qui complique la comparaison avec les termes complexes contenus dans les deux autres ressources.

Le tableau 3 montre que ces modifications entraînent, comme on pouvait s'y attendre, une réduction substantielle du nombre de couples de voisins (- 37,8 %). Le nombre de synonymes varie assez peu (peu d'unités polylexicales dans cette ressource). En revanche, l'augmentation du nombre de couples de la base JDM est patente : + 40,6 % pour la base dans son ensemble, + 91 % pour les antonymes, + 4,4 % pour les hyperonymes et + 59,2 % pour les méronymes. C'est la conséquence de la symétrisation des relations. La faiblesse du pourcentage d'augmentation des hyperonymes s'explique par le fait que l'effacement des unités complexes réduit quasiment de moitié le nombre de couples alors que cette opération n'affecte que peu les autres relations (beaucoup de couples d'hyperonymes sont composés d'un syntagme nominal et de sa tête : *bouillon/bouillon de poulet*, *bière/bière sans alcool*…).

	Avant homogénéisation				**Après homogénéisation**			
VDW	3 922 657				2 556 810			
DES	389 182				358 001			
JDM	**Ensemble**	**Anto.**	**Hypo.**	**Méro.**	**Ensemble**	**Anto.**	**Hypo.**	**Méro.**
	753 426	9946	45 515	18 646	1 059 003	18 993	45 705	29 693

Tableau 3 : Volumes des Voisins de Wikipédia (VDW), du Dictionnaire Électronique des Synonymes (DES) et de JeuxDeMots (JDM) en nombre de couples, avant et après homogénéisation.

Dans la section suivante, nous confrontons ces trois ressources, dans le but de comprendre ce qui conditionne le repérage des relations lexicales par les méthodes d'analyse distributionnelle automatique.

3 Analyser le différentiel entre voisinage distributionnel et relations lexicales

La comparaison des données est réalisée à partir du lexique commun aux ressources considérées deux à deux (VDW/DES et VDW/JDM) : nous considérons le sous-ensemble des couples du DES et de JDM dont les deux mots figurent dans le vocabulaire couvert par les voisins. Les tableaux 4 et 5 montrent que le vocabulaire des voisins se retrouve presque intégralement dans les deux ressources puisque 91 % et 96 % des couples des VDW sont constitués de deux mots apparaissant respectivement dans le DES et JDM. En revanche, l'inverse est loin d'être vrai puisque seulement 28 % des couples du DES et 27% de ceux de JDM sont constitués de deux mots apparaissant parmi les voisins. Le biais introduit par le corpus est donc important en ce qui concerne l'ampleur du vocabulaire pris en compte. Ces chiffres ne signifient cependant pas que 80% des mots n'apparaissent pas du tout dans le corpus : ils peuvent avoir des occurrences, mais trop peu nombreuses pour être prises en compte par l'analyse distributionnelle.

3.1 Comparaison globale

Le nombre de couples communs VDW/DES et VDW/JDM peut s'analyser en terme de précision et de rappel. La précision désigne la proportion de couples de voisins qui correspondent à un couple recensé dans le DES ou JDM. Dans les deux cas, elle est extrêmement basse puisqu'à peine 2 % des couples de voisins sont des synonymes (tableau 4) et environ 3 % apparaissent dans JDM (tableau 5). Cela confirme une première étude menée par Galy et Bourigault (2005). La proportion des couples d'antonymes, d'hyponymes et de méronymes parmi les voisins, détaillée dans ce deuxième tableau, est également très faible puisqu'elle se situe dans les trois cas autour de 1 %.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100293

	VDW	DES
Couples du lexique commun	2 330 212	101 597
Couples partagés	42 254	
Précision	0,02	
Rappel	0,42	

Tableau 4 : Comparaison entre les voisins de Wikipédia et le DES.

Le rappel désigne la proportion des couples figurant dans les deux ressources de référence que l'on retrouve dans les voisins. Comme on le voit dans le tableau 5, le DES et JDM sont respectivement couverts à 42 % et 29 % par les voisins. La couverture pour les sous-ensembles JDM$_{\text{ANTO}}$, JDM$_{\text{HYPO}}$ et JDM$_{\text{MÉRO}}$ varie entre 34 % et 37 %.

	VDW	JDM	VDW	JDM$_{\text{ANTO}}$	VDW	JDM$_{\text{HYPO}}$	VDW	JDM$_{\text{MÉRO}}$
Couples du lexique commun	2 466 547	288 199	231 216	4838	700 199	17 020	409 682	15 912
Couples partagés	82 533		1770		6314		5380	
Précision	0,03		0,01		0,01		0,01	
Rappel	0,29		0,37		0,37		0,34	

Tableau 5 : Comparaison entre les voisins de Wikipédia et 4 versions du réseau JDM.

Ces chiffres confirment que l'AD détecte bien autre chose que les relations lexicales habituellement recensées dans les dictionnaires. Comme nous l'avons déjà indiqué, ce n'est pas ce chiffre qui nous intéresse ici, mais le chiffre du rappel. Le premier constat que nous pouvons faire est que les quatre relations considérées sont repérées dans des proportions comparables. Néanmoins, on observe que la synonymie est un peu mieux repérée. Cette différence est difficile à interpréter, car il peut s'agir d'un effet lié à une différence de qualité entre les deux bases (JDM et le DES). On remarque également que les trois relations sur lesquelles nous nous sommes focalisés pour JDM sont mieux identifiées que la moyenne des relations couvertes par cette base. Le second constat, c'est que la majorité des paires recensées dans les deux ressources de référence ne passent pas le test de l'AD dans le corpus considéré. C'est cet aspect que nous allons explorer dans ce qui suit, en commençant par dégager quelques caractéristiques statistiques générales des « bons candidats » à l'AD.

3.2 Caractéristiques générales des voisins distributionnels

Le tableau 6 met en évidence certaines contraintes statistiques que le calcul distributionnel fait peser sur les couples de mots. Dans ce tableau, on note $A \cap B$ l'intersection de A et de B, et $A \setminus B$ la différence ensembliste de A et de B, c'est-à-dire les éléments qui appartiennent à A mais pas à B. Ces chiffres confirment que le calcul distributionnel favorise les mots qui présentent certaines propriétés en termes de fréquence et de productivité. Quelle que soit la relation considérée, la somme de la fréquence des mots qui composent les couples de voisins est plus élevée que celle des non-voisins. Il en va de même pour le critère de productivité, qui est d'ailleurs généralement corrélé à la fréquence. Les couples qui ont été repérés par l'analyse distributionnelle ont donc une productivité moyenne plus élevée que ceux qui ne l'ont pas été.

De plus, les mots qui composent les couples de voisins ont des fréquences et des productivités plus équilibrées : le rapport entre la fréquence (respectivement la productivité) des deux mots varie entre 0,42 et 0,45 pour les voisins (respectivement 0,43 à 0,46) alors qu'il oscille entre 0,2 et 0,31 pour les non-voisins (respectivement 0,18 à 0,28). Ce point est important : on en conclut que des paires de synonymes dont l'un a un sens plus spécifique que l'autre (ex : *transformer* et *déguiser*) seront moins facilement

repérées si cela se traduit par de fortes différences de fréquence dans le corpus. Bien que la mesure de Lin soit conçue pour limiter l'impact de ce déséquilibre, il reste important et doit être pris en considération.

		Fréquence	Rapport fréquence	Productivité	Rapport productivité
DES	∩ VDW	19373	0,44	575	0,43
	\ VDW	10167	0,29	284	0,25
JDM_{ANTO}	∩ VDW	13285	0,45	391	0,45
	\ VDW	8715	0,31	238	0,28
JDM_{HYPO}	∩ VDW	28060	0,42	747	0,43
	\ VDW	19897	0,2	516	0,18
$JDM_{MÉRO}$	∩ VDW	30625	0,45	797	0,46
	\ VDW	15350	0,28	416	0,25

Tableau 6 : Comparaison de la fréquence et de la productivité entre les couples de voisins et de non-voisins.

Les tendances observées dans le tableau 6 nous amènent à nous focaliser dans la suite de l'étude sur les couples de mots dont l'absence parmi les voisins n'est pas imputable à des effets liés à la fréquence ou à la productivité de leurs membres. Nous cherchons à voir pourquoi des couples de mots qui présentent des propriétés optimales pour le calcul distributionnel ne sont pourtant pas identifiés.

3.3 Étude de couples non repérés par l'analyse distributionnelle

Nous avons constitué un échantillon de couples issus du DES et de JDM et dont les propriétés sont compatibles avec l'AD, pour en faire un examen qualitatif. En d'autres termes, nous avons cherché à isoler des couples qui seraient théoriquement de bons candidats pour l'AD, et qui pourtant ne sont pas identifiés par cette méthode. Ces couples sont sélectionnés selon deux critères (N.B. nous avons opté pour l'examen de la productivité plutôt que la fréquence, les deux critères étant globalement équivalents) :

- la productivité moyenne de leurs deux membres,
- le rapport entre les productivités des deux membres.

Les paires que nous analysons en priorité sont donc celles qui ont une productivité moyenne élevée et dont les deux membres ont des productivités comparables (rapport supérieur ou égal à 0,44). Parmi celles dont les productivités moyennes étaient les plus élevées, nous avons extrait :

- les 10 premières paires de noms et de verbes pour les synonymes et les antonymes,
- les 10 premières paires de noms pour les hyperonymes et les méronymes.

3.3.1 Synonymie

Le premier tableau d'exemples (tableau 7) montre des paires de synonymes du DES qui ne présentent pas de proximité distributionnelle dans le corpus.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100293

Noms	air/aspect, mois/traitement, pied/plante, masse/public, étape/hôtel, course/distance, distance/opposition, accès/crise, approche/arrivée, croix/épreuve
Verbes	battre/tourner, monter/relever, doter/favoriser, juger/mesurer, chanter/exécuter, assister/entourer, élever/remonter, conseiller/pousser, rapporter/rattacher, aboutir/accéder

Tableau 7 : Synonymes fréquents dans le corpus mais non repérés par l'AD.

L'examen des couples de noms montre que la polysémie des termes considérés fournit une première explication à l'absence du couple parmi les voisins. Par exemple, la synonymie *air*/*aspect* porte sur une acception du mot *air* qui n'est que très peu représentée dans la distribution de ce mot dans le corpus, au profit des acceptions « fluide gazeux »[4] (que l'on trouve dans des contextes du type *air vicié*, *courant d'air*, *air refroidir*) et « mélodie » (*air d'opérette*, *danser sur l'air de*). Il en va de même pour les couples *accès/crise* et *approche/arrivée*. Si les deux mots partagent certains contextes (par exemple, *démence*, *jalousie* et *fièvre* pour le couple *accès/crise*), ceux-ci pèsent peu dans une distribution par ailleurs plus largement associée au sens spatial du mot *accès*.

Les autres couples relèvent d'un autre cas de figure. Il ne s'agit cette fois plus seulement de termes qui pourraient, dans un corpus différent, présenter une distribution plus semblable. Ils illustrent en effet des cas de synonymie extrêmement particuliers, voire douteux. Considérons les exemples *mois/traitement* ou *croix/épreuve*. La synonymie porte sur un emploi très figé d'un des mots. *Mois* a le sens de *salaire*, *traitement* principalement dans l'expression *toucher son mois*. *Croix* n'a le sens d'*épreuve* que dans le contexte *porter sa croix*. Le décalage distributionnel est donc prévisible, la non application du critère de substituabilité met au jour des couples dont l'un des termes ne peut être isolé du contexte spécifique qui justifie le rapprochement sémantique.

L'étude des couples de verbes illustre également des cas de rapprochement extrêmement spécifiques des deux mots *via* la relation de synonymie. Ainsi, *chanter* et *exécuter* ne sont proches que lorsqu'*exécuter* signifie *interpréter une chanson*. Autre exemple : *pousser* ne peut être rapproché de *conseiller* que dans des contextes très particuliers ; or, dans le corpus, les emplois de *pousser* sont très variés puisqu'il apparaît fréquemment comme verbe support (*pousser un cri*, *pousser la reconnaissance* [*jusqu'à*]), ou avec un complément d'objet concret (*pousser une porte*), ce qui exclut le rapprochement avec *conseiller*. On se rend compte que même lorsqu'il régit des compléments d'objet humains (*troupe*, *pays*, *auteur*…), il n'est pas paraphrasable par *conseiller* mais par des verbes plus génériques comme *conduire* ou *amener à* (sans restriction sur la nature du sujet). On peut noter d'ailleurs que le voisinage distributionnel repère d'autres synonymes plus proches de *conseiller*, à savoir *recommander* ou *inciter*.

L'absence de ces couples dans les voisins s'explique donc par le biais qu'introduit le corpus, lequel sélectionne des acceptions des mots qui ne correspondent pas à celle qui est visée par le couple de synonymes. Par ailleurs, cette confrontation révèle des cas de synonymie très restrictifs, où le principe de substituabilité s'applique de façon marginale, soit parce qu'on a affaire à un sens rare, soit parce que l'emploi est associé à des contextes très spécifiques, voire à du figement.

3.3.2 Antonymie

Le premier constat que l'on peut faire au vu des paires d'antonymes du tableau 8 est que ce ne sont pas, pour la plupart, des paires d'antonymes canoniques, au sens de Murphy (2003), c'est-à-dire unies par une relation d'opposition binaire conventionnelle (ex : *bonheur*/*malheur, vice*/*vertu*). Seule la paire verbale *pleurer*/*rire* relève incontestablement de cette catégorie. On voit ainsi que l'antonyme le plus approprié de *destruction* n'est pas *génération* (mais *création*), et celui d'*échec* n'est pas *résolution* (mais *réussite*, *victoire* ou *succès*). Les couples mieux assortis que sont *création*/*destruction*, *échec*/*victoire*, *échec*/*réussite*, *échec*/*succès, bonheur*/*malheur* apparaissent par contre tous dans la base de voisins. Certains couples semblent d'ailleurs contestables. C'est particulièrement le cas de *huile*/*vague* (résultant peut-être d'une généralisation abusive de l'opposition entre *mer d'huile* et *vagues* ?). On peut également s'étonner de trouver *dormir/réveiller* parmi les antonymes (la forme pronominale *se réveiller* serait plus adéquate) ; de même pour *interrompre*/*progresser*.

DOI 10.1051/shsconf/20120100293

Noms	destruction/génération, accident/substance, échec/résolution, huile/vague, franchise/obligation, adhésion/refus, bonheur/douleur, défaut/vertu, déclin/enfance, assurance/peur
Verbes	bâtir/renverser, interrompre/progresser, décliner/progresser, anéantir/fortifier, agiter/calmer, embaucher/virer, chiffrer/déchiffrer, dormir/réveiller, attiser/modérer, pleurer/rire

Tableau 8 : Antonymes non repérés par l'analyse distributionnelle.

Comme dans le cas de la synonymie, on constate ensuite que l'on a affaire à des antonymes partiels dont le sens concerné par la relation est minoritaire dans le corpus. C'est le cas de *vertu*/*défaut* (il est plutôt question dans le corpus de *vertu* au sens de propriété) ou *déclin*/*enfance* (*déclin* ne désigne généralement pas un processus affectant l'individu). Le même phénomène s'observe sur les verbes, notamment dans le cas de *bâtir*/*renverser* : *bâtir* prend pour objets des noms de bâtiments ainsi que quelques noms abstraits, dont certains sont partagés avec *renverser* (*empire*, *royaume*), mais celui-ci s'emploie principalement au sens figuré de « provoquer la chute de, venir à bout de, anéantir », et privilégie en position objet des noms désignant des régimes politiques comme *monarchie*, *république*, *empire*, *dictature* ou les individus qui les représentent (*roi*, *prince*, *président*, *empereur*, *dictateur*…).

On constate enfin que certains de ces couples d'antonymes, par exemple *résolution*/*conflit*, ou *bonheur*/*douleur*, ont précisément la particularité de s'associer à des contextes de nature très différente. Ainsi, si les mots *résolution* et *conflit* renvoient tous deux à des évènements, la résolution porte sur des situations conflictuelles (*conflit*, *différend*, *crise*, *paradoxe*…), ce qui n'est pas le cas de l'échec (*attaque*, *expédition*, *projet*, *révolte*…). Les mots *bonheur* et *douleur* désignent certes des sentiments que l'on éprouve, mais chacun se spécialise dans une gamme de contextes bien distincte. En d'autres termes, dans le cas de ces antonymes, le principe d'opposition se traduit par une divergence sur le plan distributionnel.

3.3.3 Hyperonymie et méronymie

Le dernier tableau concerne seulement des couples de noms, et présente conjointement des cas de méronymie et d'hyperonymie.

Hyperonymes	lieu/championnat, direction/sud, oiseau/pape, orchestre/rock, endroit/tribunal, fleur/pensée, poète/racine, organe/yeux, juge/métier, métier/réalisateur
Méronymes	arrivée/circuit, bijou/coffre, pédale/roulement, crabe/pince, prince/tête, orgue/registre, foyer/incendie, aigle/patte, plastique/tuyau, chat/queue

Tableau 9 : Hyperonymes et méronymes non repérés par l'AD.

On observe tout d'abord dans la liste d'hyperonymes deux couples problématiques car ils ne passent pas le test proposé par Cruse pour identifier cette relation, à savoir la possibilité pour le couple Y/X d'intégrer le patron *X est un type de Y.* S'il est vrai par exemple que le sud est une direction et le pape un oiseau, le rock n'est pas un orchestre, ni le championnat un lieu. Le couple *poète/racine* pose, lui, un problème particulier d'homonymie nom commun/nom propre. Les sept couples restants illustrent des cas de relation d'hyperonymie dans lesquels les deux termes ne sont pas substituables dans le corpus. Trois d'entre eux posent à nouveau des problèmes de polysémie (*pape*, *pensée*, *organe*). Restent quatre couples (*direction*/*sud*, *endroit*/*tribunal*, *juge*/*métier*, *réalisateur*/*métier*) qui présentent un véritable décalage distributionnel. Ils illustrent tous les quatre le fait que le terme spécifique n'est plus conçu en contexte comme une instance du terme générique – il n'hérite pas de son type sémantique : *juge* et *réalisateur* entrent dans des contextes désignant des individus, *tribunal* dans des contextes désignant un collectif humain, *sud* désigne une zone et non une direction. L'étude de l'hyperonymie par le biais de l'AD offre un point de vue intéressant sur la question de la catégorisation sémantique qui opère effectivement dans le discours.

Le cas de la méronymie est particulier. Le lien entre cette relation et le principe de substituabilité ne va pas de soi, si l'on considère la diversité des sous-types de relation qui sont couverts par la méronymie (Winston *et al.* 1987). On peut par exemple s'attendre à ce que les contextes partagés par un nom

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100293

désignant un artefact et un nom désignant son composant soient très limités. Néanmoins, on a vu qu'un tiers des méronymes de JDM étaient identifiés par l'AD, proportion à peine moins importante que celle des relations précédentes, ce qui nous amène à la considérer au même titre. La liste de méronymes que nous présentons dans le tableau 9 illustre deux types de méronymie : composant/objet dans un cas (*plastique*/*tuyau*), constituant/objet dans tous les autres (*pince*/ *crabe*, *arrivée*/*circuit,* etc.). Le décalage distributionnel entre les deux membres du couple semble cette fois évident. Si l'on prend par exemple le cas des parties du corps, représentées dans quatre couples (*pince*/*crabe*, *patte*/*aigle*, *queue*/*chat*, *tête*/*prince*), il est clair que le rapprochement n'est possible qu'à condition que le tout soit considéré sous son angle anatomique, ce qui n'est que très marginalement le cas dans le corpus : ainsi, le mot *chat* apparaît principalement dans des contextes adjectivaux (*domestique*, *errant*, *sauvage*…) parmi lesquels seuls quelques adjectifs de couleur seraient attribuables à ses parties du corps. Dans le cas de la méronymie, il semble donc plus intéressant de se demander sous quelles conditions le principe de substituabilité s'applique – et, par exemple, quels types de méronymie sont les plus susceptibles d'y répondre. L'examen rapide des couples de méronymes qui sont également des voisins semble par exemple montrer une prédominance de la relation membre/collection (*bateau*/*flotte*, *musicien*/*orchestre*) bien que la relation composant/objet puisse également figurer (*farine*/*céréale*, *eau*/*corps*). Nous avons consacré une étude plus systématique de ce phénomène (à paraître) qui confirme notamment que les couples de méronymes de type membre/collection sont particulièrement bien repérés par l'AD. Cela est dû au fait que leurs deux membres peuvent apparaître dans des contextes similaires (*naviguer*_SUJ, *couler*_OBJ ou encore *équipage_de* pour le couple *bateau/flotte*).

L'observation d'un petit échantillon de couples pour les quatre relations étudiées permet de dégager différentes explications possibles aux limites du test de substituabilité. La polysémie en est une. Les couples qui ne répondent pas au test de substituabilité illustrent alors des cas de relation partielle : le sens représenté dans la relation n'est que marginalement représenté dans le corpus. Si le corpus est suffisamment vaste et diversifié, ces décalages peuvent alors être révélateurs de paires synonymiques correspondant à des acceptions marginales. Nous avons vu par ailleurs que ce test permettait de repérer d'autres sources de décalage entre la ressource de référence et les propriétés distributionnelles des mots dans le corpus : emplois figés dans le cas de la synonymie, antonymes non canoniques voire douteux, hyponymes dont la catégorisation sémantique s'émancipe en discours de celle de leur terme générique. Cette première approche, de nature exploratoire, suggère donc des pistes pour étudier de façon plus systématique ces causes de décalage.

3.4 Étude des différences de couverture entre mots

La deuxième méthode d'observation des données que nous avons choisie est ici mise en œuvre sur la relation de synonymie uniquement. Elle consiste à partir cette fois des mots qui apparaissent dans les voisins et à leur appliquer les mesures de précision, de rappel (cf. 3.1), ainsi que la mesure F qui les combine[5] (Manning et Schütze, 1999). L'utilisation de ces critères nous permet d'observer sous deux angles différents les propriétés distributionnelles des mots du corpus. Ainsi, pour le sous-ensemble des synonymes :

- la précision pour un mot donné est le rapport entre le nombre de ses voisins qui apparaissent parmi ses synonymes et le nombre de ceux qui n'y apparaissent pas,
- le rappel est le rapport entre le nombre de ses synonymes repérés par les voisins et le nombre total de ses synonymes dans le DES.

Le calcul de la mesure F nous permet de prendre en compte ces deux aspects de la distribution des mots : pour qu'un mot ait une mesure F élevée, il faut que ses voisins couvrent la plus grande proportion de couples recensés pour ce mot dans le dictionnaire des synonymes tout en produisant un minimum de *bruit*, c'est-à-dire de paires de voisins n'y apparaissant pas. Nous nous appuyons sur ces mesures pour

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100293

faire émerger les mots dont les voisins ne recoupent que très peu (voire pas du tout) les données du DES dans le but de mettre au jour leurs caractéristiques.

3.4.1 Propriétés générales des voisins en terme de précision/rappel

Nous considérons trois versions de la base des voisins dans le tableau 10, en faisant varier la valeur de la mesure de Lin, de manière à observer le comportement des voisins selon le degré de proximité distributionnelle considéré. Nous distinguons cette fois les voisins selon leur catégorie grammaticale.

Seuil (Lin)	Nombre de couples	Catégorie	Nombre de voisins	Nombre de synonymes	Nombre de voisins synonymes	Précision	Rappel	Mesure F
0,1	2 330 212	Ensemble	278	12	5	0,05	0,35	0,05
		Noms	306	12	5	0,04	0,36	0,05
		Verbes	366	18	8	0,05	0,37	0,05
		Adjectifs	86	9	3	0,09	0,32	0,08
0,2	300 477	Ensemble	42	12	2	0,10	0,18	0,08
		Noms	39	11	2	0,09	0,17	0,08
		Verbes	78	18	4	0,08	0,18	0,07
		Adjectifs	14	7	1	0,17	0,18	0,11
0,3	45 747	Ensemble	9	10	1	0,18	0,11	0,10
		Noms	7	9	1	0,17	0,10	0,10
		Verbes	18	16	2	0,14	0,10	0,08
		Adjectifs	4	5	1	0,27	0,16	0,16

Tableau 10 : Comparaison des propriétés de trois versions des VDW.

L'écart important que l'on peut observer entre la précision et le rappel pour la base seuillée à 0,1 est dû au fait que le nombre de voisins extraits pour un mot est toujours largement supérieur au nombre de ses synonymes dans le DES. Cela a pour conséquence de favoriser le rappel au détriment de la précision. Le nombre de voisins par mot chutant considérablement avec l'augmentation du seuil, ce phénomène est un peu moins marqué dans la version seuillée à 0,2 et s'inverse dans la version à 0,3. On remarque que ce sont les adjectifs qui ont la précision la plus élevée. Cela est-il simplement dû au fait que les adjectifs sont la catégorie qui a le moins de voisins ? Il semblerait que non : le tableau 11 montre qu'à nombre de voisins équivalent, la catégorie des adjectifs reste celle qui a la meilleure précision. Ce phénomène est d'autant plus remarquable que, comme le montre le tableau 10, les adjectifs sont les mots pour lesquels le DES compte le moins de synonymes. La différence de précision entre les noms, verbes et adjectifs tend toutefois à s'estomper avec l'augmentation du nombre de voisins.

		Noms	Verbes	Adjectifs
Nombre de voisins	de 1 à 5	0,11	0,13	0,2
	de 6 à 10	0,07	0,09	0,11
	de 11 à 15	0,06	0,06	0,08

Tableau 11: comparaison de la précision des noms, verbes et adjectifs à nombre de voisins équivalents.

Pour les trois versions de la base, la mesure F reste très basse mais l'on peut observer une légère augmentation : la hausse de la précision a plus d'influence que la baisse du rappel. Toutefois, alors que la mesure F de la base seuillée à 0,3 est deux fois plus élevée que pour celle à 0,1, le nombre moyen de voisins synonymes par mot est divisé par cinq. Cela signifie que le seuillage de la base implique un compromis entre rappel et précision avec d'un côté, une ressource très bruitée couvrant une grande proportion des synonymes, et de l'autre, une ressource qui contient une plus grande proportion de synonymes mais dont la couverture est considérablement réduite.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100293

3.4.2 Analyse des propriétés des mots en fonction de leur mesure F

Nous analysons deux types de mots dans ce qui suit : ceux dont les voisins recouvrent les synonymes dans des proportions importantes et ceux pour lesquels ce n'est pas le cas. Nous cherchons à comprendre ce qui conditionne cette différence de comportement vis-à-vis de l'AD.

Nous nous appuyons sur la mesure F pour différencier ces deux ensembles. Afin d'éviter de prendre en compte les couples qui ne sont pas détectés par l'AD à cause de leur différence de productivité (cf. 3.2), nous avons choisi d'écarter tous ceux dont le rapport de productivité était inférieur à la moyenne (0,33). La base obtenue compte 6727 mots : 4102 noms, 1401 verbes et 1224 adjectifs.

3.4.2.1 Mots dont la mesure F est élevée

Le tri des mots par mesure F décroissante fait émerger des mots qui, pour la plupart, ont très peu de voisins et très peu de synonymes : le nombre de voisins moyen des 34 mots qui ont une mesure F supérieure ou égale à 0,5 est de 2,6 et leur nombre moyen de synonymes est de 2,1. Certains mots se distinguent par un nombre un peu plus élevé de voisins. Le tableau 12 rapporte les 10 mots ayant la mesure F la plus élevée une fois les mots ayant moins de 10 voisins et moins de 10 synonymes écartés. On constate que 9 de ces 10 mots sont des adjectifs. Cette proportion confirme la tendance – observée à la section 3.4.1 – qu'ont les adjectifs à avoir une précision élevée, mais elle reste remarquable étant donné que les adjectifs ne constituent que 18 % des mots de notre liste. Il reste toutefois difficile de dire si ces résultats sont révélateurs d'un fonctionnement spécifique des adjectifs dans notre corpus (d'autant que les adjectifs qui émergent ont la particularité d'exprimer une appréciation du scripteur, alors que la subjectivité est – théoriquement – bannie de Wikipédia) ou s'ils reflètent une propriété générale de la relation modifieur, qui générerait moins de bruit dans les voisins qu'elle permet de rapprocher que les autres relations exploitées lors de l'AD.

Mot	Catégorie	Nombre de voisins	Nombre de synonymes	Nombre de voisins synonymes	Précision	Rappel	Mesure F
étonnant	A	55	29	17	0,31	0,59	0,40
colossal	A	16	14	6	0,38	0,43	0,40
prodigieux	A	17	20	7	0,41	0,35	0,38
fabuleux	A	23	28	9	0,39	0,32	0,35
formidable	A	21	17	6	0,29	0,35	0,32
honorable	A	16	10	4	0,25	0,40	0,31
terrible	A	57	22	12	0,21	0,55	0,30
merveilleux	A	64	22	13	0,20	0,59	0,30
zèle	N	17	10	4	0,24	0,40	0,30
incroyable	A	38	17	8	0,21	0,47	0,29

Tableau 12 : Les dix mots ayant la mesure F la plus élevée.

3.4.2.2 Mots dont la mesure F est nulle

La différence entre le nombre moyen de voisins et de synonymes par mot implique que la plupart des mots ont une mesure F extrêmement basse (cf. 3.4.1). Ainsi, un mot comme *intéresser* a un nombre tellement élevé de voisins – 1221 – que sa mesure F est de 0,04 alors que l'ensemble de ses 20 synonymes a été capté (le rappel est de 1, la précision de 0,02). Nous nous intéressons ici aux mots dont aucun des synonymes n'a pu être capté par l'AD (leur précision, rappel et mesure F est donc de 0). Comme dans la section précédente, nous n'avons pas pris en compte les mots ayant moins de 10 voisins. Le tableau 13 rapporte, pour chaque catégorie, quelques-uns des mots parmi ceux qui ont le plus de synonymes (sous la forme *mot* (*nombre de voisins*/*nombre de synonymes*)).

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100293

Noms	agrément (33/19), cœur (17/18), flamme (80/14), conformité (20/14), excitation (18/14), commencement (102/13), illusion (50/13), accompagnement (64/12), touche (55/12)
Verbes	arranger (45/26), parer (233/16), grouper (12/16), atténuer (13/14), épuiser (16/13), adjoindre (33/12), allonger (22/12), capter (50/11), ajuster (40/11), enrichir (299/10)
Adjectifs	rude (15/24), grossier (12/18), timide (18/13), barbare (71/11), valable (19/10), tendre (12/10), rustique (10/10), nuisible (13/9), digne (40/8), spontané (20/8)

Tableau 13 : Exemples de mots pour lesquels aucun des synonymes n'a été capté.

On peut principalement distinguer deux raisons expliquant le décalage entre les synonymes d'un mot et ses voisins distributionnels. La plus évidente est celle de la polysémie, déjà évoquée dans les analyses de la section 3.3 : les sens du mot *excitation* qui émergent du corpus Wikipédia correspondent principalement à ses acceptions du point de vue physiologique (voisins : *infection*, *lésion*, *pathologie*…) et de celui de la physique (voisins : *ionisation*, *vibration*, *radiation*…), ce qui implique que l'ensemble de ses synonymes relevant du domaine des états mentaux (*effervescence, encouragement, enthousiasme…*) ne sont pas détectés. Ainsi, *excitation* n'apparaît pas dans des contextes comme *susciter*_OBJ ou *exprimer*_OBJ, comme c'est le cas pour plusieurs de ses synonymes, mais plutôt dans des SN comme *lumière d'excitation*, *courant d'excitation* ou *spectre d'excitation*, qui apparaissent comme des termes appartenant à des domaines de spécialité. On peut également citer le cas de l'adjectif *rustique*, dont les emplois dans le corpus relèvent de l'acception « résistant, robuste, qui demande peu de soin ». Ce sens est absent des synonymes, qui se répartissent entre les acceptions « relatif à l'agriculture, à la vie des champs » (*champêtre*, *pastoral*, *paysan*, *rural*), « sans savoir-vivre » (*grossier*, *rude*, *sauvage*, *vulgaire*) et « sans apprêt, brut » (*brut*, *primitif*, *simple*). Nous avons ici affaire à des situations où des acceptions entières d'un mot ne sont pas détectées par l'analyse du corpus : ce décalage est révélateur d'emplois atypiques du mot, au regard de la représentation qu'en donne le dictionnaire.

Certains décalages entre voisins et synonymes s'apparentant à de la polysémie peuvent s'expliquer par des différences de registre ou des emplois figurés. Le registre du corpus peut en effet exclure l'emploi de certains synonymes : la plupart des synonymes du nom *touche* ne sont pas repérés car ils renvoient à l'acception familière « aspect général d'une personne, d'une chose » du mot (*allure, apparence, look, maintien…*), alors que l'analyse fait émerger le sens de « commande manuelle » (voisins : *bouton, dispositif, clavier*…). C'est également le cas de *veine*, dont les synonymes liés à son acception « chance, fortune » (*hasard, prospérité, réussite, bonheur, pot…*) ne sont pas détectés par l'analyse (ses voisins – *artère, nerf, muscle*… – sont ici aussi liés à son sens premier). De même, les synonymes relevant d'emplois figurés sont peu repérés : ainsi, le nom *commencement* possède de nombreux synonymes métaphoriques relevant des champs sémantiques de la croissance humaine (*adolescence*, *berceau*, *embryon*, *enfance*, *naissance*) ou végétale (*éclosion*, *fleur*, *germe*, *racine*). Le cas de *flamme* est également emblématique à ce titre puisque beaucoup de ses synonymes renvoient à des sentiments (*désir, passion, élan, enthousiasme*…) alors que ce n'est pas le cas de ses voisins.

Une deuxième raison pouvant expliquer l'absence de certains synonymes d'un mot parmi ses voisins est le fait que même s'ils partagent un noyau de sens identique, ils ne se manifestent pas dans des contextes similaires. Ainsi, le verbe *atténuer* partage le même noyau de sens que ses synonymes *abaisser*, *affaiblir*, *alléger* et *apaiser*, mais ces derniers n'apparaissent pas parmi ses voisins. La raison en est que ces quatre verbes sélectionnent des types d'objets différents de ceux du verbe *atténuer* (lequel s'emploie principalement avec des noms renvoyant à des phénomènes physiques comme *vibration, bruit, fréquence, son*…) : *abaisser* porte sur des noms de mesure (*prix*, *taux seuil*, *niveau*…), *affaiblir* sur des noms désignant des humains (*adversaire*, *roi*), des ensembles d'humains (*armée*, *population*) ou des organisations (*régime*, *parti*), *alléger* sur des noms exprimant une notion de poids (*masse*, *charge*, ou, au sens figuré, *souffrance*), *apaiser* sur des phénomènes (*querelle*, *crise*) ou des états mentaux (*colère*, *mécontentement*).

Ainsi, dans le cas de *veine*, *touche*, *commencement* et *flamme*, les synonymes ne sont pas repérés par l'AD parce que l'acception sur laquelle porte la relation de synonymie ne se manifeste pas – ou trop peu –

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100293

dans le corpus. Dans le cas de mots comme *atténuer*, les synonymes partagent un même sens mais se distinguent du point de vue de leurs distributions : le critère de la substituabilité se trouve mis à mal, dans la mesure où les contextes sélectionnés par le mot et ses synonymes sont tellement différents que l'analyse ne permet pas de les rapprocher.

4 Conclusion

L'utilisation combinée d'une base distributionnelle et de deux ressources externes permet de mettre à l'épreuve le critère de substituabilité, considéré habituellement comme un test pour apprécier la propension de deux mots à entretenir une relation lexicale. Cette étude montre la difficulté à appréhender le contenu d'une base distributionnelle pléthorique, résultat de l'analyse d'un vaste corpus de textes caractérisé par l'hétérogénéité des termes abordés. La confrontation avec des lexiques déjà constitués fournit un angle d'étude réducteur mais éclairant. Elle montre que la proximité sémantique mise au jour par l'analyse distributionnelle dépasse très largement celle dont ces lexiques rendent compte. Elle montre également que les relations lexicales ne se traduisent pas systématiquement par une proximité distributionnelle effective dans un vaste corpus. En nous concentrant sur ce deuxième aspect, nous avons dégagé différents éléments d'analyse expliquant ce décalage. Les premiers éléments sont liés au mode de calcul de l'AD, qui favorise la mise au jour de relations entre des mots non seulement fréquents dans le corpus, mais de fréquence comparable. Cela ajoute une contrainte statistique forte sur le test de substituabilité. Ceci étant posé, nous avons montré que l'importance du décalage entre les voisins et les deux autres ressources lexicales utilisées fournissait un angle d'étude intéressant sur l'opposition entre des relations attribuées *in abstracto* et des relations construites dans le discours. Ainsi, certaines relations d'hyperonymie ne sont pas opérantes dans le corpus, car la catégorisation sémantique qu'elles induisent n'est pas mobilisée dans le texte ; des mots n'ont aucun synonyme parmi leurs voisins parce que leur acception dans le corpus n'est pas prise en compte dans le dictionnaire ; des rapports de synonymie très spécifiques, relevant d'emplois restrictifs voire figés, ne se traduisent par aucune proximité distributionnelle. L'utilisation de l'AD en appoint de la construction de ressources lexicales génériques peut ainsi permettre d'introduire des informations relatives au caractère central ou marginal de la relation dans différents corpus. Cette étude suggère d'autres pistes d'analyse, relatives à la différence de comportement des relations lexicales vis-à-vis du test de substituabilité, comme on a pu le voir dans le cas de l'antonymie (certains antonymes semblent avoir une distribution nettement disjointe) ou de la méronymie (seuls certains types de méronymie semblent se prêter au test de substitution). L'analyse distributionnelle automatique fournit donc un observatoire intéressant pour étudier de façon empirique la manifestation des relations sémantiques en discours.

Références bibliographiques

Agirre E., Alfonseca E., Hall K., Kravalova J. et Soroa A. (2009). A study on similarity and relatedness using distributional and WordNet-based approaches. *Proceedings of NAACL-HLT*.

Baroni M. et Lenci A. (2010). Distributional Memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):1-49.

Baroni M. et Lenci A. (2011). How we BLESSed distributional semantic evaluation. *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*.

Bourigault D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *Actes de la 9e conférence sur le Traitement Automatique de la Langue Naturelle*, 75-84.

Bourigault D. (2007). Un analyseur syntaxique opérationnel : SYNTEX. *Mémoire d'habilitation à diriger des recherches*. Université Toulouse II – Le Mirail.

Bouaud J., Habert B., Nazarenko A. et Zweigenbaum P. (2000). Regroupements issus de dépendances syntaxiques sur un corpus de spécialité : catégorisation et confrontation à deux conceptualisations du domaine. *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, Charlet J., Zacklad M., Kassel G. et Bourigault D. (eds), Eyrolles, Paris, 275-290.

Cruse D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Dias G., Moraliyski R., Cordeiro J. et Doucet A. (2010). Automatic discovery of word semantic relations using paraphrase alignment and distributional lexical semantics analysis. *Natural Language Engineering*, 1(1):1-30.

Galy E. et Bourigault D. (2005). Analyse distributionnelle de corpus de langue générale et synonymie. *4es Journées de la linguistique de corpus*, 2005.

Grefenstette G. (1992). SEXTANT: exploring unexplored contexts for semantic extraction from syntactic analysis. *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 324-326.

Grefenstette G. (1994). Corpus-derived first-, second- and third-order word affinities. *Proceedings of Euralex,* Amsterdam, 279-290.

Harris Z. (1954). Distributional structure. *Word,* 10(23):146–162.

Hearst M. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th International Conference on Computational Linguistics*. Association for Computational Linguistics, 539-545.

Hindle D. (1990). Noun classification from predicate-argument structure. *Proceedings of the 28th annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 268-275.

Kilgarriff A. et Yallop C. (2000). What's in a thesaurus?. *Proceedings of the Second Conference on Language Resources and Evaluation*, 1371-1379.

Lafourcade M. (2007). Making people play for lexical acquisition. *Proceedings of the 7th Symposium on Natural Language Processing*.

Lin D. (1998). An information-theoretic definition of similarity. *Proceedings of the 15th International Conference on Machine Learning*, 296-304.

Manguin J.-L., François J., Eufe R., Fesenmeier L., Ouzouf C. et Sénéchal M. (2004). Le dictionnaire électronique des synonymes du CRISCO : un mode d'emploi à trois niveaux. *Cahiers du CRISCO* 17, Université de Caen.

Manning C. D. et Schütze H. (1999). *Foundations of statistical natural language processing*, MIT Press, Cambridge.

Murphy L. (2003). *Semantic relations and the lexicon*. Cambridge University Press.

Nazarenko A., Zweigenbaum P., Bouaud J. et Habert B. (1997). Corpus-Based Identification and Refinement of Semantic Classes. *Proceedings of the 1997 American Medical Informatics Association* (AMIA). AMIA.585-589.

Ruge G. (1992). Experiments on linguistically-based term associations. Information Processing and Management, 28(3):317-332.

Sahlgren M., (2006). Towards pertinent evaluation methodologies for word-space models. *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

Turney P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. *COLING*, 905-912.

van der Plas L. (2008). Automatic lexico-semantic acquisition for question answering. Thèse de doctorat, Université de Groningen (Pays-bas).

Winston M., Chaffin R. et Herrmann D. (1987). A taxonomy of part-whole relations, *Cognitive Science*, 11: 417-441.

---

[1] La ressource est consultable à l'adresse suivante : http://redac.univ-tlse2.fr/applications/vdw.html

[2] Site du DES : http://www.crisco.unicaen.fr/des/

[3] Site de JeuxDeMots : http://www.lirmm.fr/jeuxdemots/

[4] Ces définitions sont extraites du Trésor de la Langue française : http://atilf.atilf.fr/tlf.htm

[5] *F = 2*(précision*rappel)/(précision+rappel)*